



Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



CovidPhy: A tool for phylogeographic analysis of SARS-CoV-2 variation

Xabier Bello^{a,b}, Jacobo Pardo-Seco^{a,b}, Alberto Gómez-Carballa^{a,b}, Hansi Weissensteiner^c, Federico Martín-Torres^{a,d}, Antonio Salas^{a,b,*}

^a Genetics, Vaccines and Pediatric Infectious Diseases Research Group (GENVIP), Instituto de Investigación Sanitaria de Santiago (IDIS) and Universidad de Santiago de Compostela (USC), Galicia, Spain

^b Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain

^c Institute of Genetic Epidemiology, Department of Genetics and Pharmacology, Medical University of Innsbruck, 6020, Innsbruck, Austria

^d Translational Pediatrics and Infectious Diseases, Department of Pediatrics, Hospital Clínico Universitario de Santiago de Compostela, Galicia, Spain

ARTICLE INFO

Keywords:

SARS-CoV-2

COVID-19

RNA

Superspreading events

Variants of concern

Phylogeny

ABSTRACT

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the pathogen responsible for the coronavirus disease 2019 (COVID-19) pandemic. SARS-CoV-2 genomes have been sequenced massively and worldwide and are now available in different public genome repositories. There is much interest in generating bioinformatic tools capable to analyze and interpret SARS-CoV-2 variation. We have designed CovidPhy (<http://covidphy.eu>), a web interface that can process SARS-CoV-2 genome sequences in plain fasta text format or provided through identity codes from the Global Initiative on Sharing Avian Influenza Data (GISAID) or GenBank. CovidPhy aggregates information available on the large GISAID database (>1.49 M genomes). Sequences are first aligned against the reference sequence and the interface provides different sources of information, including automatic classification of genomes into a pre-computed phylogeny and phylogeographic information, haplogroup/lineage frequencies, and sequencing variation, indicating also if the genome contains known variants of concern (VOC). Additionally, CovidPhy allows searching for variants and haplotypes introduced by the user and includes a list of genomes that are good candidates for being responsible for large outbreaks worldwide, most likely mediated by important superspreading events, indicating their possible geographic epicenters and their relative impact as recorded in the GISAID database.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a single-stranded RNA virus responsible for the coronavirus disease 2019 (COVID-19) pandemic. There has been a massive interest in sequencing genomes from coronavirus circulating in COVID-19 patients worldwide since its first early sequencing in December 2019 (Wu et al., 2020). Genomes are stored in public repositories such as GenBank (<https://www.ncbi.nlm.nih.gov/sars-cov-2/>) or, more specifically, in The Global Initiative on Sharing Avian Influenza Data (GISAID; <https://www.gisaid.org>) (Shu and McCauley, 2017) as well as the 2019 Novel Coronavirus Resource (2019nCoV; <https://bigd.big.ac.cn/ncov/online/tools>).

During the last few months, several software applications and web tools have been developed that aim at understanding SARS-CoV-2

variation as well as dissemination in a worldwide scale. One of the most popular tool is Nextstrain (<https://nextstrain.org> (Hadfield et al., 2018)), which provides a maximum likelihood phylogeny built on a massive amount of SARS-CoV-2 genomes, and which allows to investigate the phylodynamics of the virus since the beginning of the pandemic. However, nomenclature of the Nextstrain phylogeny is limited to a few nodes among hundreds, it does not follow systematic criteria for naming clades, and it is not stable since it has undergone several changes over the last few months. Moreover, mutational pathways along branches from the root to a given node can only be reconstructed partially in most of the tree; therefore, although very informative from the phylodynamics point of view, the phylogeny does not allow to classify a genome into a phylogenetic node with the exception of a few nodes of interest (see, in contrast, human mtDNA phylogeny including 5,500 haplogroups (van Oven, 2015) and related

* Corresponding author. Unidade de Xenética, Instituto de Ciencias Forenses (INCIFOR), Facultade de Medicina, Universidade de Santiago de Compostela, and GenPoB Research Group, Instituto de Investigación Sanitaria (IDIS), Hospital Clínico Universitario de Santiago (SERGAS), Galicia, Spain.

E-mail address: antonio.salas@usc.es (A. Salas).

<https://doi.org/10.1016/j.envres.2021.111909>

Received 25 July 2021; Received in revised form 16 August 2021; Accepted 17 August 2021

Available online 20 August 2021

0013-9351/© 2021 Elsevier Inc. All rights reserved.

classification tools such as Haplogrep (Weissensteiner et al., 2016). The tool Nextclade, which is part of the Nextstrain project, allows to classify SARS-CoV-2 fasta sequences into 18 major clades. Besides mutation calling and clade assignment, it performs some quality checks and phylogenetic placement in a global tree. The largest SARS-CoV-2 clade dataset is provided by the Pango Network comprising 1,801 virus lineages (out of which 514 lineages are withdrawn) with its own lineage naming rules (Rambaut et al., 2020). Several tools for online and offline classification, lineage suggestion with an own lineage designation committee, tree modification and reporting are provided. Other analytical tools for the analysis of SARS-CoV-2 variation include CovSeq, a Python and JavaScript designed web interface (Liu et al., 2020a) that aggregates data from different repositories to extract information on genetic variants that, ultimately, can be downloaded by users. The varSEAK database (<https://varseak.bio>) offers a variety of analyses, providing information on variants, lineages, and a splice site prediction tool. Other software includes Viral Genome ORF Reader (VIGOR (Wang et al., 2010)), which focuses on gene annotation, VAPiD (Shean et al., 2019), a pipeline that facilitates genome submissions to NCBI GenBank, and the National Genomics Data Center (Gong et al., 2020), an online tool that includes BLAST alignment, genome annotation, variant identification modules, among others utilities.

We have developed CovidPhy (www.covidphy.eu), a web tool that allows to process and analyze complete SARS-CoV-2 genomes. CovidPhy implements a pipeline that accepts newly generated sequencing fasta files, but also the identification codes of genomes stored in GISAID or GenBank repositories. It classifies genomes into main phylogenetic nodes and offers information on viral variants and clade frequencies worldwide.

By inspecting the large GISAID database, it makes it possible to identify specific SARS-CoV-2 sequences as strong candidates for being responsible for notable COVID-19 outbreaks. The first attempt was carried out by Gómez-Carballa et al. (2020a); in this early article, we explored the database available at that time (containing >4.7 K SARS-CoV-2 genomes) for identical genomes that showed a high frequency in a short time frame (of only a few days) and occurring in specific geographic areas. This signature, coupled with the substitution rate of the SARS-CoV-2 (which generates a substitution approximately every two weeks) and the incubation period needed for the development of the COVID-19 symptoms (5–6 days, according to WHO on November 8, 2020) signaled the likely presence of sudden local outbreaks that could have been originated by superspreading events. Topological inspection of the phylogenies originated by these candidates added further support to a model of germ transmission compatible with superspreading and not with alternative ways of transmission (e.g. chains). This procedure was subsequently extended in Gómez-Carballa et al. (2020b) to explore the important outbreaks occurring in Spain during the first wave of the pandemic, which was particularly devastating in this country. Further investigations corroborated the important role of superspreading in the COVID-19 pandemic (Adam et al., 2020; Althouse et al., 2020; Liu et al., 2020b; Walker et al., 2020). The article by Lemieux et al. (2020) specifically treated two important outbreaks occurring in Boston in the early weeks of the pandemic; notably, one of these event had been recorded by our early analysis in Gómez-Carballa et al. (2020a); this feature was pointed out in Salas et al. (2021). Therefore, the identification of genomes that were responsible for important outbreaks in different countries and locations by simply inspecting the GISAID database constitutes another useful feature of CovidPhy.

2. Material and methods

2.1. Data source and processing

CovidPhy uses data from GISAID to compute variant and clade (haplogroup) frequencies and infer SARS-CoV-2 candidates responsible

for outbreaks. Genomes are aligned against the reference genome with GenBank accession number MN908947.3 (submitted on January 5, 2020) corresponding to the first SARS-CoV-2 genome released on GenBank (GISAID ID #402125). Most of the sequences used by CovidPhy were incrementally downloaded from GISAID since our initial publications and aligned to the reference sequence as previously indicated (Gómez-Carballa et al. 2020a, b; Pardo-Seco et al., 2021; Salas et al., 2021). A total of 1,493,746 genomes (downloaded on February 5, 2021 from GISAID) are now being processed in CovidPhy. We extracted the differences between any sequence and the reference, and the genomes were classified into the nodes of a pre-generated phylogeny.

2.2. Phylogeny and nomenclature

We implemented the phylogeny (and nomenclature) built by Gómez-Carballa et al. (2020a, b) to classify SARS-CoV-2 genomes into clades; this is also available and navigable in the web interface. Although built on data produced during the first wave of the pandemic, to the best of our knowledge this phylogeny remains the most elaborate one available. The lack of a consensus nomenclature has generated great controversy among the scientific community (<https://www.newscientist.com/article/mg24933242-900-coronavirus-variant-names-are-too-confusing-there-is-a-better-way/> (Callaway, 2021)), as also echoed by the media (e.g. <https://www.nytimes.com/2021/03/02/health/virus-variant-names.html>), especially with the identification of new variants of concern (VOC) in late December 2020. We had already warned about this controversy in our early publication (Gómez-Carballa et al., 2020a). Thereby, the differing naming for SARS-CoV-2 genetic lineages by GISAID, Nextstrain and Pango will remain by the scientific community, so that it seems most reasonable to keep the consistent nomenclature employed in the source mentioned. In addition, note also that the minimal nomenclature scheme employed by Nextstrain does not follows a systematic criteria for naming branches and the nomenclature is dynamic (see comments in (Gómez-Carballa et al., 2020a)); this is also a common feature of the nomenclature used by Rambaut et al. (2020, 2021). Instead, as of June 2021, The World Health Organization (WHO) proposed a new naming convention for VOCs and variants of interest (<https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>) introducing letters of the Greek alphabet; CovidPhy provides classification of sequences into the four popular VOCs, namely, Alpha, Beta, Gamma and Delta. Together with this classification scheme, CovidPhy also identifies worrying mutations according to the CoVariants resource (<https://covariants.org/shared-mutations>), many of these mutations are relevant to VOCs.

2.3. Software stack

The whole stack was written in Nim programming language (<https://nim-lang.org>). Nim is one of the best performing languages (<https://github.com/kostya/benchmarks>; <https://github.com/def/nim-benchmarksgame>), usually on par with C, without sacrificing readability and expressiveness.

The alignment of the input sequences is performed with a modified version of MAFFT (Katoh and Standley, 2013), so it can be interfaced directly through a foreign function interface (FFI). The database of choice is SQLite (<https://www.sqlite.org/>), as we deemed that the strengths of that database (easy installation and management, capable of handling web traffic over 500 K hits/day) outweighed the weaknesses (not ready for high concurrency writes, no client/server structure). Nim can also compile to JavaScript, targeting the browser and allowing the developer to write the full stack of a web service in one single language.

The graphics for the webpage are created using Plotly (<https://plotly.com/javascript/>), interfaced using Nim both in the frontend and in the backend. The web stack is a single binary, but the core (the aligner and the classifier) is decoupled, and therefore it could be reused to build also a graphical user interface (GUI) and command line interface (CLI)

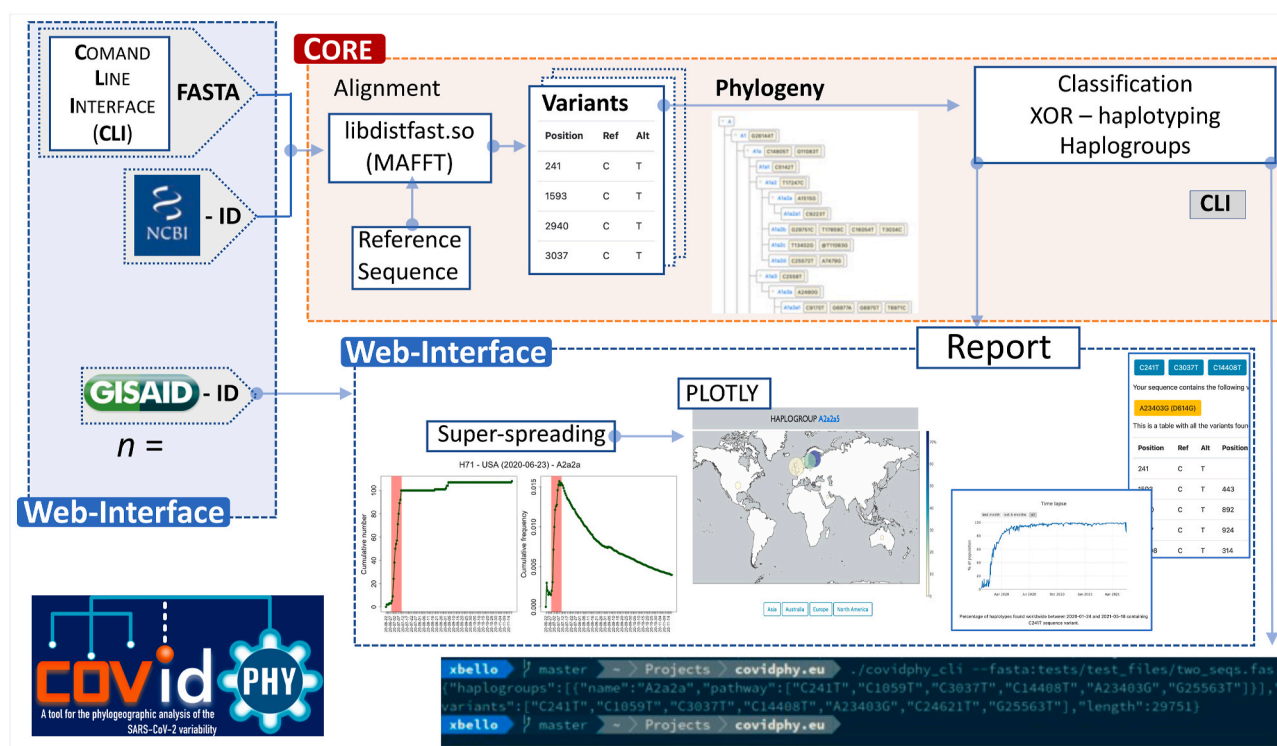


Fig. 1. Pipeline of CovidPhy. CovidPhy offers three interfaces: a web, a CLI and a GUI. All three can be fed with a fasta file (top left) that is aligned using *libdistfast.so* against the Reference (402,125) and scanned looking for differences that allow the classification in a precomputed phylogeny (top, red square marked “core”). The output varies for each program: the CLI and the GUI only output the haplogroup and the variants found (bottom black square), while the web offers additional information: haplogroup frequencies in regions (e.g. countries), candidates for important outbreaks as inferred from database searchers, and VOCs. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

(Fig. 1). We provide three main programs in the repository:

- covidphy, the web server that can be reached at www.covidphy.eu;
- covidphy_cli, a command line interface that is used to classify sequences in our internal pipelines;
- covidphy_gui, a simple graphic interface for those unfamiliar with CLI, who may prefer to select the input with buttons

2.4. Detection of outbreak candidates

Outbreak candidates were investigated in GISAID by searching identical haplotypes detected at least 30 times in a period of five consecutive days in a specific country or state (note however that these values might change in future updates of the tool depending on e.g. database size). Once an outbreak is detected, we determine its length by adding consecutive days after the identified event, so long as the displaced 5-day window meets the previous condition of at least 30 identical sequences; alternatively, we reduce the interval to the extreme if the number of counts equals to 0 while still satisfying the minimum criterion of at least 30 identical haplotypes in the shortened period. These analyses were carried out using R software (R core Team, 2019).

3. Results

3.1. Analysis carried out by CovidPhy

CovidPhy admits fasta files and can also run genomes provided through a GISAID or a GenBank identification code. The sequences are aligned against the SARS-CoV-2 reference sequence MN908947.3. New genomes are then investigated individually by exploring the variants present, their frequency information as extracted from the large GISAID database, and their phylogenetic allocation with regards to the reference

phylogeny implemented in CovidPhy.

It takes about 0.1 s to align a single genome and carry out the clade classification; the rest of the analysis is even faster because the results displayed have been precomputed.

While sequence codes from NCBI can be retrieved automatically from the website, GISAID does not allow sharing genomes stored in their database; therefore, to investigate these sequences, the user must first register in the GISAID platform and download them the fasta files directly. When a GISAID code is entered, CovidPhy only provides information on its lineage/haplogroup assignment and frequency, but not on the variants involved.

For the genomes uploaded directly in the fasta text box or those indicated as a NCBI code, the user also obtains information on the variants present against the reference sequence (MN908947.3). This basic information includes the nucleotide position, the reference variant, and the alternative allele, the ORF assignment, the predicted severity of the variant for the virus (low: if synonymous; medium: if non-synonymous; high: if it leads to stop gain/loss, frameshift, or start loss), their functional description (missense, synonymous, etc.), and its frequency in the database. In addition, the user is informed if the sequence carries worrying mutations, indicating the nucleotide and the amino acid change (if any; e.g. A23063T (N501Y) that is characteristic of the B.1.1.7 and other VOC (Davies et al., 2021)). Information on the length and coverage of the genomes is also provided (note that the 5' and 3' ends of the genomes stored in GISAID are usually missed).

Genomes are automatically classified into a clade according to the phylogeny provided by Gómez-Carballa et al. (2020a, b) and also indicates if the sequence is classified in any of the four most important VOC, but the algorithm can be applied to any classification tree. Once a sequence is classified into a clade, information is provided on the geographic location of this clade by way of displaying a continental map and including information on continental or country haplogroup

frequencies. The phylogeny used by CovidPhy is provided in full in a tab. Briefly, there are two main clades that are phylogenetically located at the same level. Although there is not a clear consensus on the root of the tree (see a discussion in (Gómez-Carballa et al., 2020a), the nomenclature is based on classical cladistics and, for practical purposes, it assumes the root in haplogroup A. Haplogroup names are organized and named hierarchically from this root: A > A1 > A1a > A1a2, and the branch variants are indicated. By clicking on an haplogroup label, the user is moved to another tab indicating geographic frequencies of this lineage/haplogroup as well as the diagnostic mutational path that characterizes it.

Finally, another notable feature of CovidPhy is to search genomes carrying a particular variants or a set of variants in the GISAID database and selecting by country. For a specific query, the tool informs on the frequency of all the haplotypes in the database containing the variant(s) in the query through the time since the beginning of the pandemic.

3.2. Outbreaks recorded in SARS-CoV-2 databases

By inspecting the large database of GISAID, it is possible to identify specific SARS-CoV-2 sequences as good candidates for being responsible for large sudden COVID-19 outbreaks triggered by superspreading events. These events are listed in CovidPhy by geographic region and by country. Information is provided on the number of sequences represented in the database for the exponential growth period, and the relative frequency of the responsible genome against the other genomes circulating in the same region during the same timeframe.

Lineage assignation of the candidate genome is provided, and the continental frequency of this haplogroup can also be graphically displayed on a map.

4. Discussion

The COVID-19 pandemic has impacted every region of the world. There is much interest in investigating and tracking evolutionary characteristics of SARS-CoV-2 variants and lineages. Apart from the interest of CovidPhy for research, there is also demand from e.g. microbiological units in hospitals lacking the bioinformatic tools for the treatment of the genome sequences that they generate on a daily basis. There are similar tools available that can process SARS-CoV-2 genome sequences and carry out different kinds of analyses. Compared to previous developments, CovidPhy offers additional features. For instance, it automatically classifies a given genome into a clade by providing a SARS-CoV-2 phylogeny, and to provide phylogeographic information for this genome. Additionally, it allows variant(s) searches in the large GISAID database providing with information on frequencies for haplotypes containing the queried variation. It also provides information on lineages that have played a critical role in the dispersal of the SARS-CoV-2 pathogen, by initiating rapid and sudden outbreaks across the world.

CovidPhy has been specifically designed for treating information on the SARS-CoV-2 sequences, but it can be easily scaled to other micro-organisms of interest for which large datasets are available. (e.g. for influenza in GISAID).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge GISAID and contributing laboratories for giving us access to the SAR-CoV-2 genome database. This study received support from projects: GePEM (Instituto de Salud Carlos III(ISCIII)/PI16/01478/Cofinanciado FEDER), DIAVIR (Instituto de Salud Carlos III

(ISCIII)/DTS19/00049/Cofinanciado FEDER; Proyecto de Desarrollo Tecnológico en Salud), Resvi-Omics (Instituto de Salud Carlos III (ISCIII)/PI19/01039/Cofinanciado FEDER), BI-BACVIR (PRIS-3; Agencia de Conocimiento en Salud (ACIS)—Servicio Gallego de Salud (SERGAS)—Xunta de Galicia; Spain), Programa Traslaciona Covid-19 (ACIS—Servicio Gallego de Salud (SERGAS)—Xunta de Galicia; Spain) and Axencia Galega de Innovación (GAIN; IN607B 2020/08—Xunta de Galicia; Spain) awarded to A.S.; and projects ReSVinext (Instituto de Salud Carlos III (ISCIII)/PI16/01569/Cofinanciado FEDER), Enterogen (Instituto de Salud Carlos III(ISCIII)/PI19/01090/Cofinanciado FEDER), and Axencia Galega de Innovación (GAIN; IN845D 2020/23—Xunta de Galicia; Spain) awarded to F.M.-T.

References

- Adam, D.C., Wu, P., Wong, J.Y., Lau, E.H.Y., Tsang, T.K., Cauchemez, S., Leung, G.M., Cowling, B.J., 2020. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.* 26, 1714–1719.
- Althouse, B.M., Wenger, E.A., Miller, J.C., Scarpino, S.V., Allard, A., Hebert-Dufresne, L., Hu, H., 2020. Superspreading events in the transmission dynamics of SARS-CoV-2: opportunities for interventions and control. *PLoS Biol.* 18, e3000897.
- Callaway, E., 2021. 'A bloody mess': confusion reigns over naming of new COVID variants. *Nature* 589, 339.
- Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A.B., Russell, T.W., Tully, D.C., Washburne, A.D., et al., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 372.
- Gómez-Carballa, A., Bello, X., Pardo-Seco, J., Martínón-Torres, F., Salas, A., 2020a. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. *Genome Res.* 30, 1434–1448.
- Gómez-Carballa, A., Bello, X., Pardo-Seco, J., Pérez Del Molino, M.L., Martínón-Torres, F., Salas, A., 2020b. Phylogeography of SARS-CoV-2 pandemic in Spain: a story of multiple introductions, micro-geographic stratification, founder effects, and super-spreaders. *Zool. Res.* <https://doi.org/10.24272/j.issn.2095-8137.2020.217-1-16>.
- Gong, Z., Zhu, J.W., Li, C.P., Jiang, S., Ma, L.N., Tang, B.X., Zou, D., Chen, M.L., Sun, Y. B., Song, S.H., et al., 2020. An online coronavirus analysis platform from the National Genomics Data Center. *Zool. Res.* 41, 705–708.
- Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123.
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
- Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C.H., Krashnikov, L.A., et al., 2020. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science*. <https://doi.org/10.1126/science.abe3261>.
- Liu, B., Liu, K., Zhang, H., Zhang, L., Bian, Y., Huang, L., 2020a. CoV-seq, a new tool for SARS-CoV-2 genome analysis and visualization: development and usability study. *J. Med. Internet Res.* 22, e22299.
- Liu, Y., Eggo, R.M., Kucharski, A.J., 2020b. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet* 395, e47.
- Pardo-Seco, J., Gomez-Carballa, A., Bello, X., Martinon-Torres, F., Salas, A., 2021. Pitfalls of barcodes in the study of worldwide SARS-CoV-2 variation and phylodynamics. *Zool. Res.* <https://doi.org/10.24272/j.issn.2095-8137.2020.364-1-7>.
- R core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rambaut, A., Holmes, E.C., O'Toole, A., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407.
- Rambaut, A., Holmes, E.C., O'Toole, A., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2021. Addendum: a dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 6, 415.
- Salas, A., Bello, X., Pardo-Seco, J., Martínón-Torres, F., Gómez-Carballa, A., 2021. Superspreading: the engine of the SARS-CoV-2 pandemic. *Science*. <https://doi.org/10.1126/science.abe3261>.
- Shean, R.C., Makhous, N., Stoddard, G.D., Lin, M.J., Greninger, A.L., 2019. VAPiD: a lightweight cross-platform viral annotation pipeline and identification tool to facilitate virus genome submissions to NCBI GenBank. *BMC Bioinf.* 20, 48.
- Shu, Y., McCauley, J., 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.* 22.
- van Oven, M., 2015. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. Int.: Genet. Suppl. Series S*, e392–e394. <https://doi.org/10.1016/j.fsigss.2015.09.155>.
- Walker, A., Houwaart, T., Wienemann, T., Vasconcelos, M.K., Strelow, D., Senff, T., Hulse, L., Adams, O., Andree, M., Hauka, S., et al., 2020. Genetic structure of SARS-CoV-2 reflects clonal superspreading and multiple independent introduction events, North-Rhine Westphalia, Germany, February and March 2020. *Euro Surveill.* 25.

- Wang, S., Sundaram, J.P., Spiro, D., 2010. VIGOR, an annotation program for small viral genomes. *BMC Bioinf.* 11, 451.
- Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., Schonherr, S., 2016. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63.
- Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al., 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.